



US009087125B2

(12) **United States Patent**  
**Berg et al.**

(10) **Patent No.:** **US 9,087,125 B2**  
(45) **Date of Patent:** **Jul. 21, 2015**

(54) **ROBUST VIDEO RETRIEVAL UTILIZING VIDEO DATA**

(56) **References Cited**

U.S. PATENT DOCUMENTS

(75) Inventors: **Jared S. Berg**, Cambridge, MA (US);  
**Andrew E. Merlino, Jr.**, Burlington,  
MA (US); **Dean R. Doody**, Middletown,  
MD (US)

(73) Assignee: **Pixel Forensics, Inc.**, Burlington, MA  
(US)

(\*) Notice: Subject to any disclaimer, the term of this  
patent is extended or adjusted under 35  
U.S.C. 154(b) by 688 days.

(21) Appl. No.: **12/853,008**

(22) Filed: **Aug. 9, 2010**

(65) **Prior Publication Data**

US 2011/0085734 A1 Apr. 14, 2011

**Related U.S. Application Data**

(63) Continuation-in-part of application No. 12/538,476,  
filed on Aug. 10, 2009.

(51) **Int. Cl.**  
**G06K 9/46** (2006.01)  
**G06F 17/30** (2006.01)

(52) **U.S. Cl.**  
CPC .... **G06F 17/30802** (2013.01); **G06F 17/30787**  
(2013.01)

(58) **Field of Classification Search**  
CPC ..... G06F 17/30784; G06F 17/30825;  
G06F 17/30802; G06F 17/30787; H04N  
17/004; H04N 17/00; G06K 9/00711; G06K  
9/00758; G06K 9/00744; Y10S 707/99933  
USPC ..... 382/195; 707/723, E17.014  
See application file for complete search history.

5,434,933 A \* 7/1995 Karnin et al. .... 382/317  
5,982,979 A 11/1999 Omata et al.  
6,404,925 B1 6/2002 Foote et al.  
6,496,221 B1 \* 12/2002 Wolf et al. .... 348/192  
6,675,174 B1 1/2004 Bolle et al.  
6,751,360 B1 \* 6/2004 Lu ..... 382/278  
7,167,574 B2 1/2007 Kim  
7,486,827 B2 2/2009 Kim  
7,532,804 B2 \* 5/2009 Kim ..... 386/252  
8,077,363 B2 12/2011 Sasaki  
8,145,656 B2 3/2012 Shatz et al.  
2003/0200217 A1 \* 10/2003 Ackerman ..... 707/9  
2004/0258397 A1 12/2004 Kim  
2009/0040311 A1 \* 2/2009 Okamoto et al. .... 348/181  
2009/0102859 A1 \* 4/2009 Athsani et al. .... 345/619  
2009/0175538 A1 \* 7/2009 Bronstein et al. .... 382/173  
2010/0306193 A1 \* 12/2010 Pereira et al. .... 707/728

OTHER PUBLICATIONS

Charles W. Therrien Defining Correlation Functions and Power Spectra for Multirate Random Process Publication 2005 pp. 6010-6013.\*

Hauptmann, A.G., et al., "Video Retrieval with Multiple Image Search Strategies," *Joint Conference on Digital Libraries (JCDL '02)*, pp. 376, Portland, OR., (Jul. 13-17, 2002).

Chen, L. and Stentiford, F.W.M., "Video sequence matching based on temporal ordinal measurement," *Pattern Recognition Letters*, 29(13):1824-1831 (Oct. 1, 2008).

Chen, L., et al., "An integrated approach to video retrieval, ACM International Conference Proceedings Series vol. 313," *Proceedings of the nineteenth conference on Australasian database*, vol. 75: 49-55(2008).

\* cited by examiner

Primary Examiner — Kim Vu

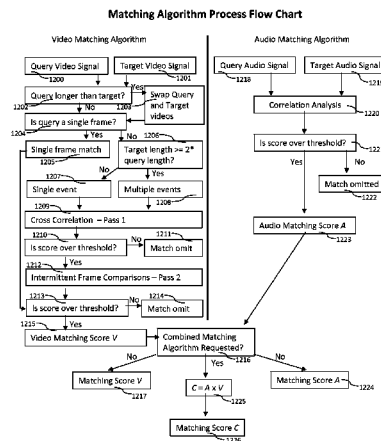
Assistant Examiner — Michael Vanchy, Jr.

(74) Attorney, Agent, or Firm — Cesari and McKenna, LLP

(57) **ABSTRACT**

Techniques for determining if two video signals match by extracting features from a first and second video signal, and cross-correlating the features thereby providing a cross-correlation score at each of a number of time lags, and finally determining the similarity score based on both the cross-correlation scores.

**18 Claims, 16 Drawing Sheets**



Video Feature Extraction Process Flow Chart

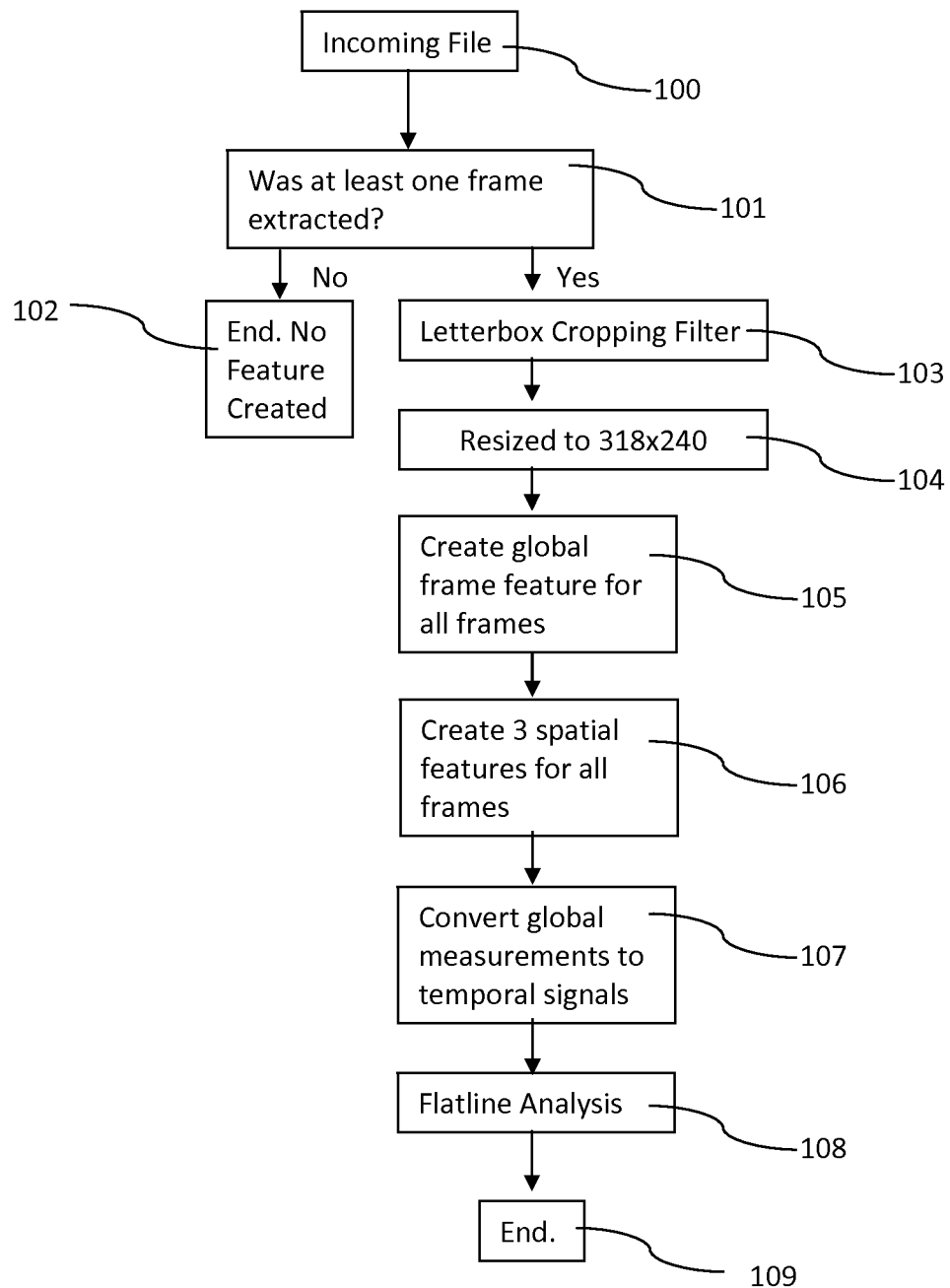


Fig. 1

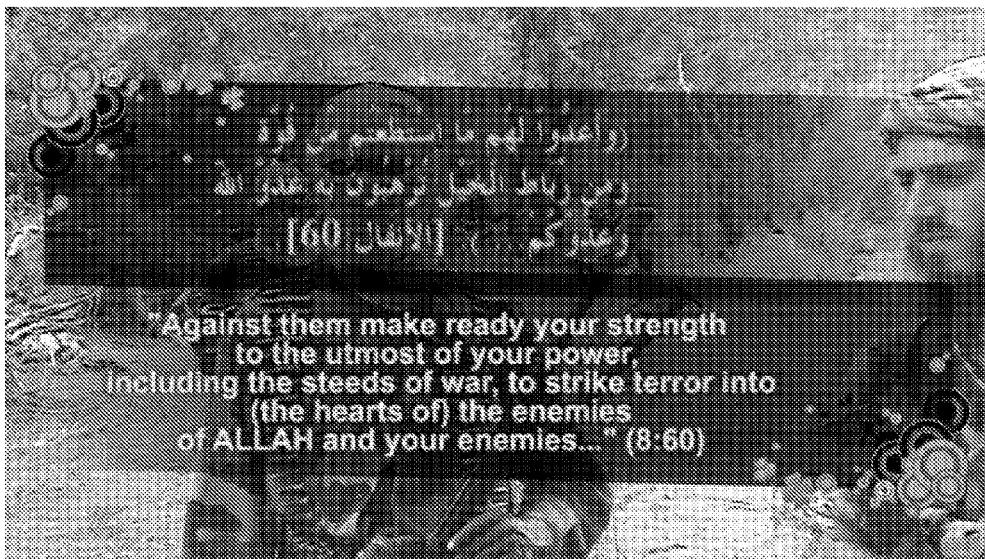


Fig. 2A



Fig. 2B

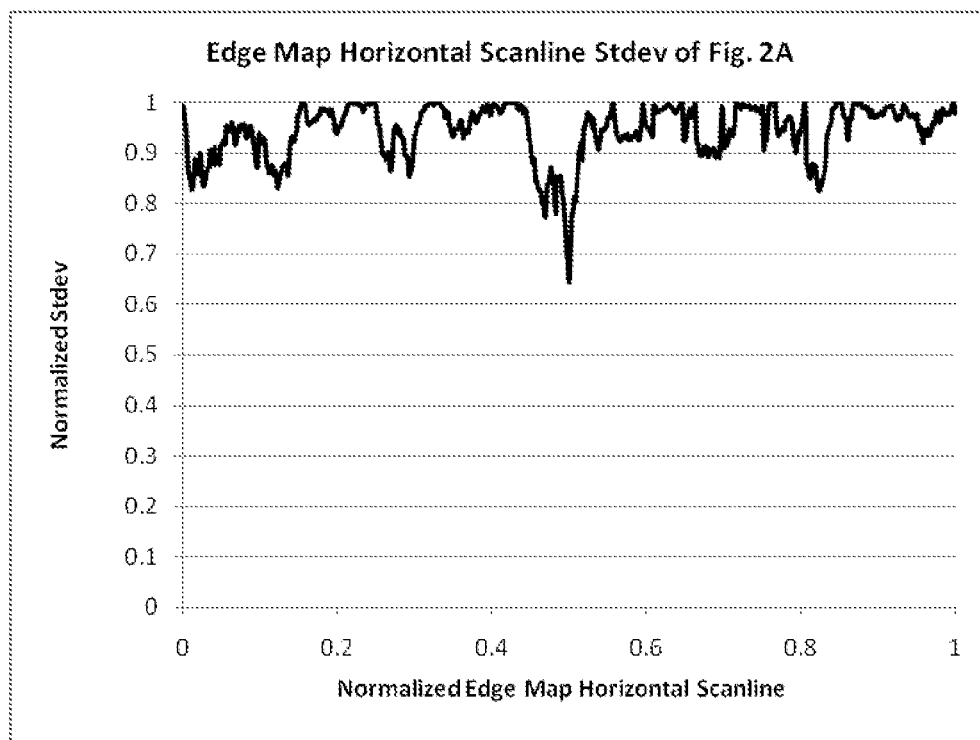


Fig. 3A

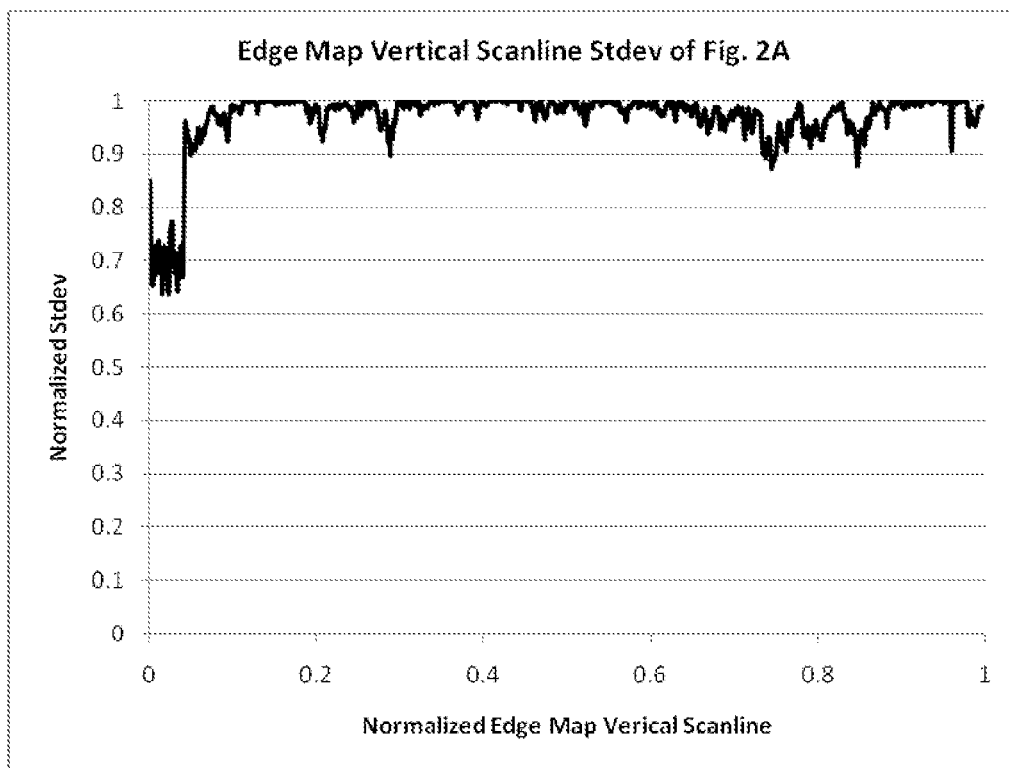


Fig. 3B

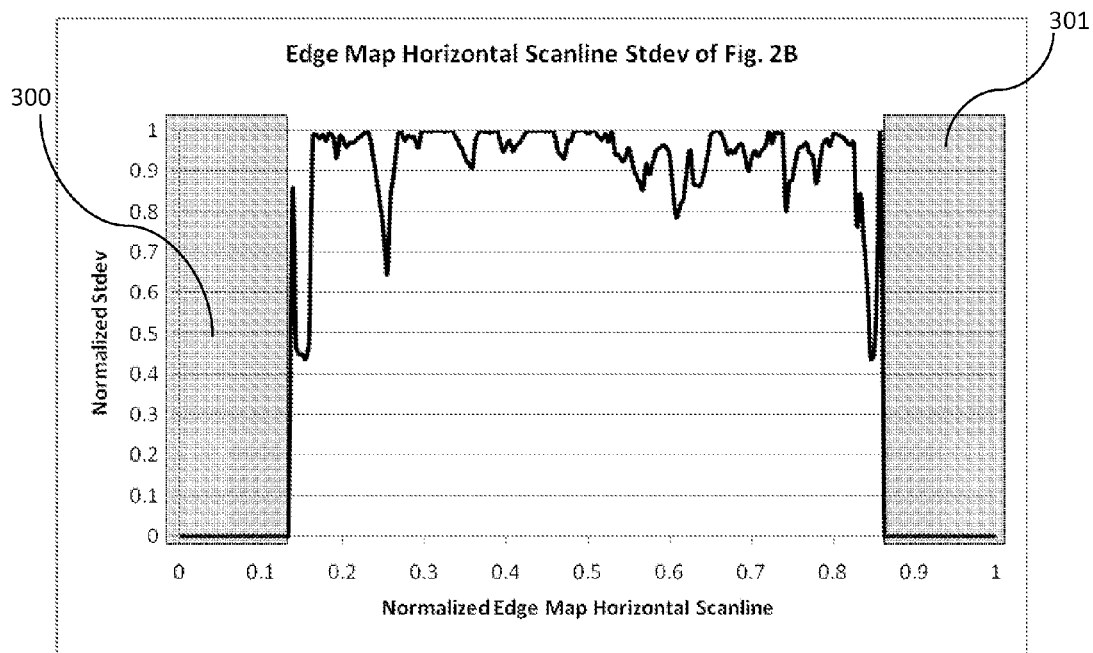


Fig. 3C

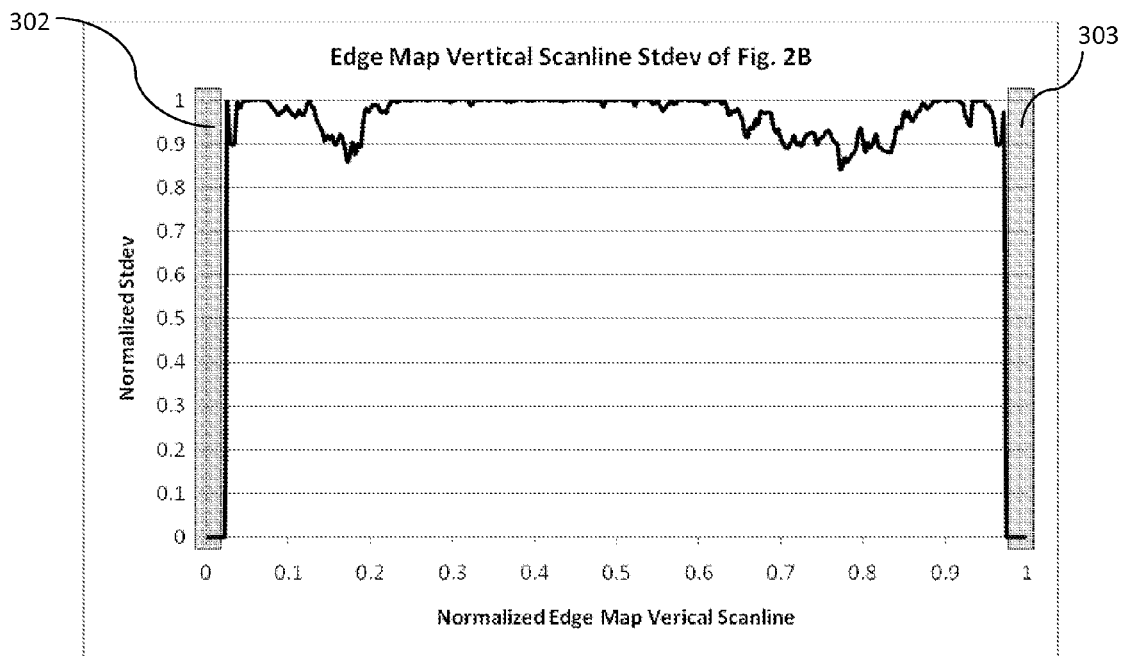


Fig. 3D

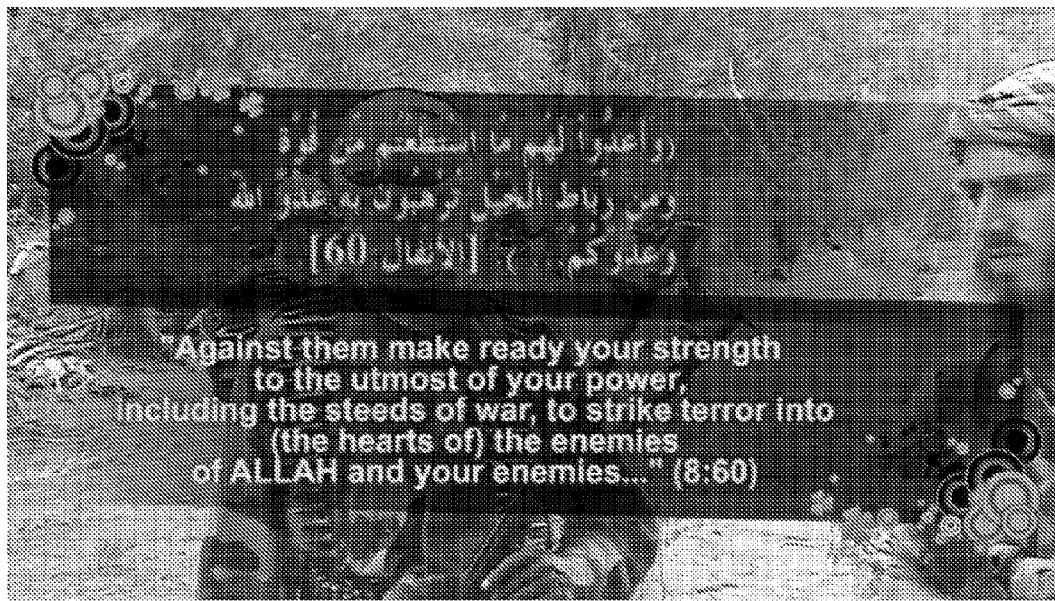


Fig. 4A

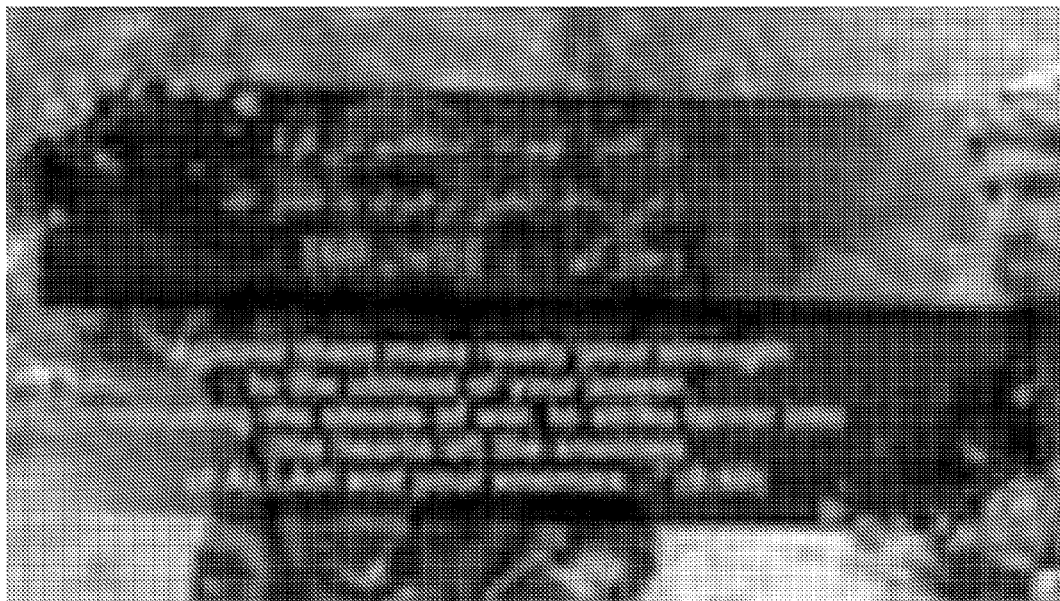
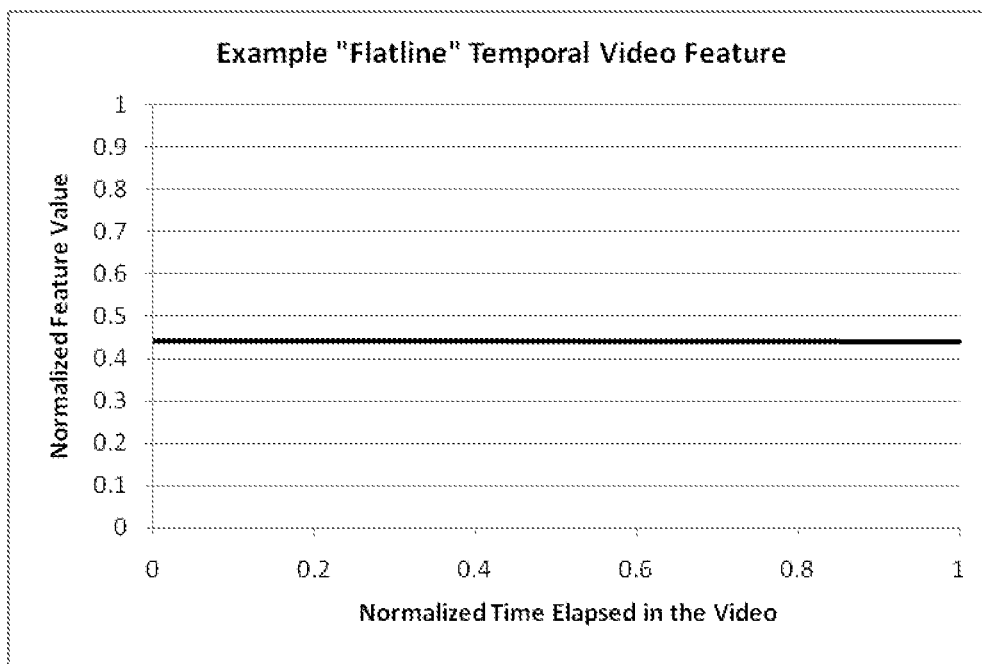
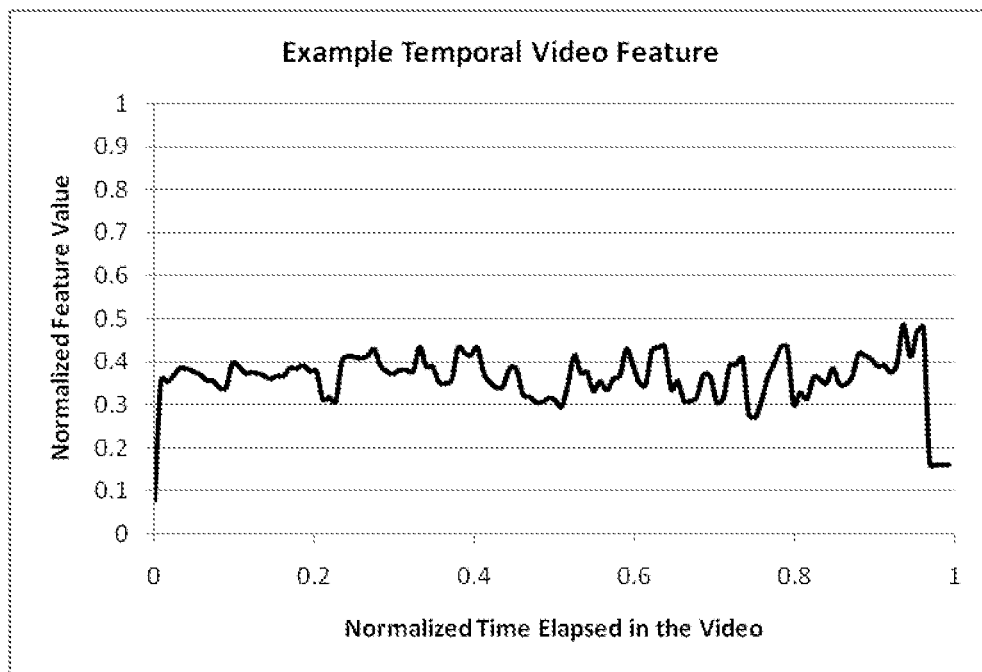


Fig. 4B



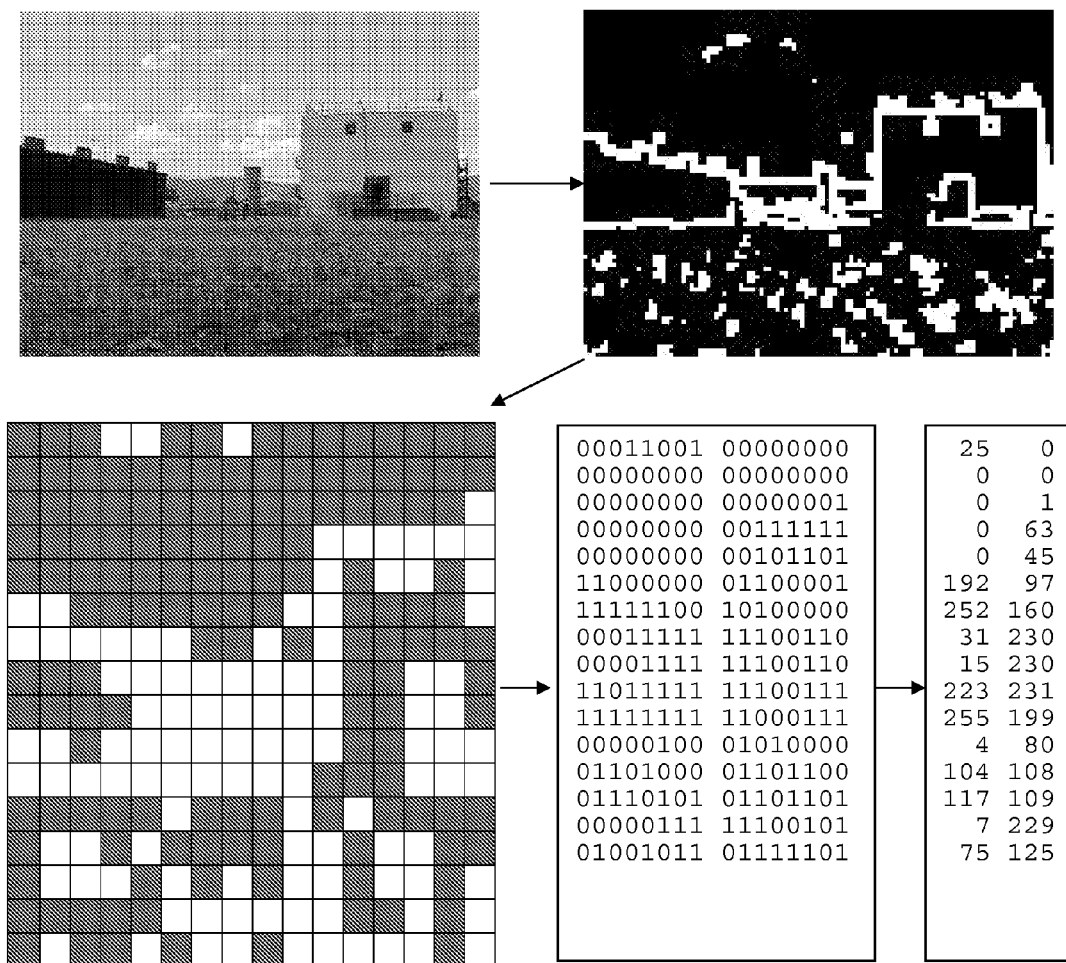


Fig. 7



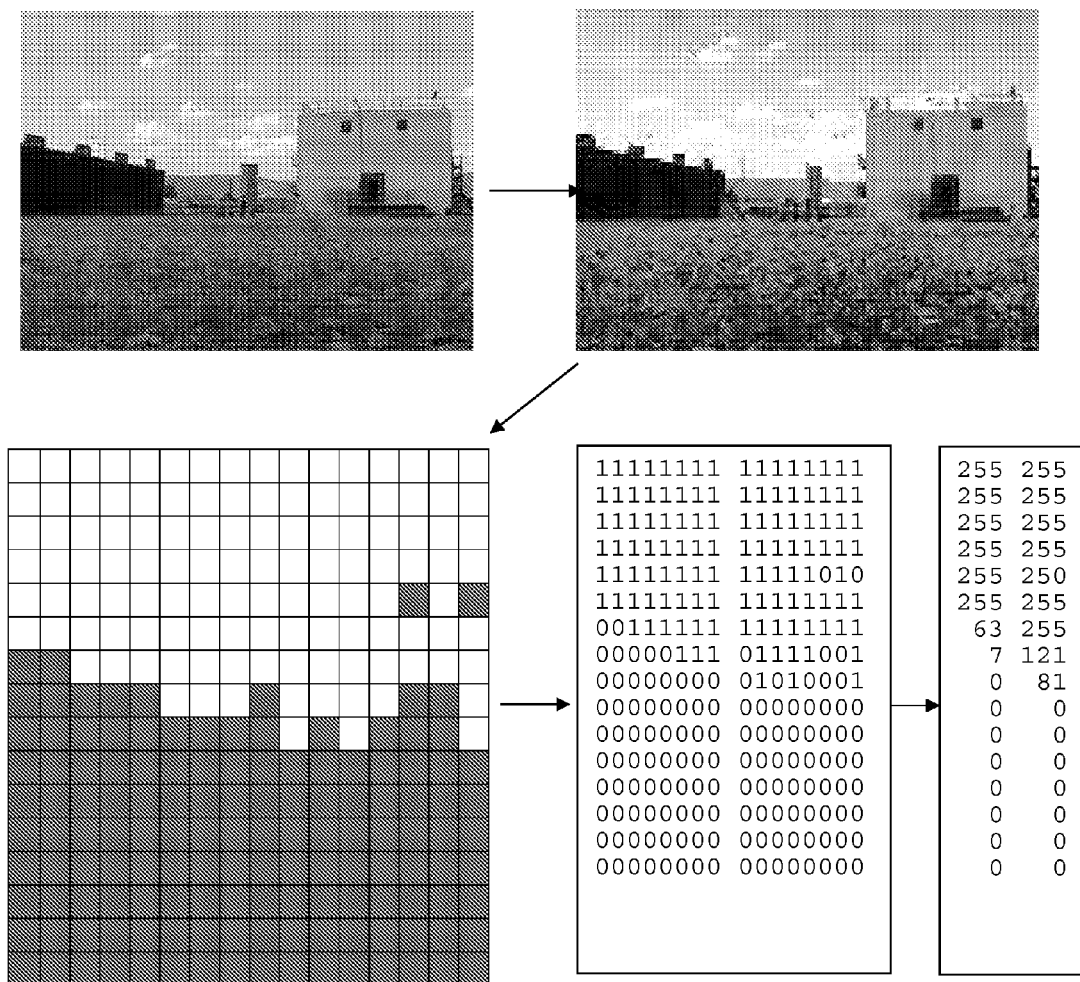
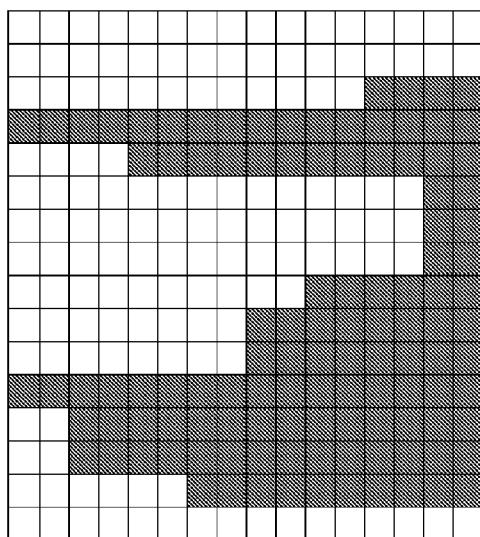
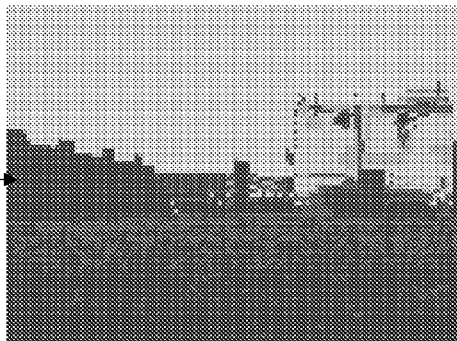
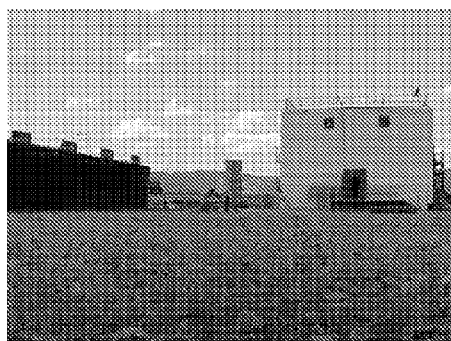


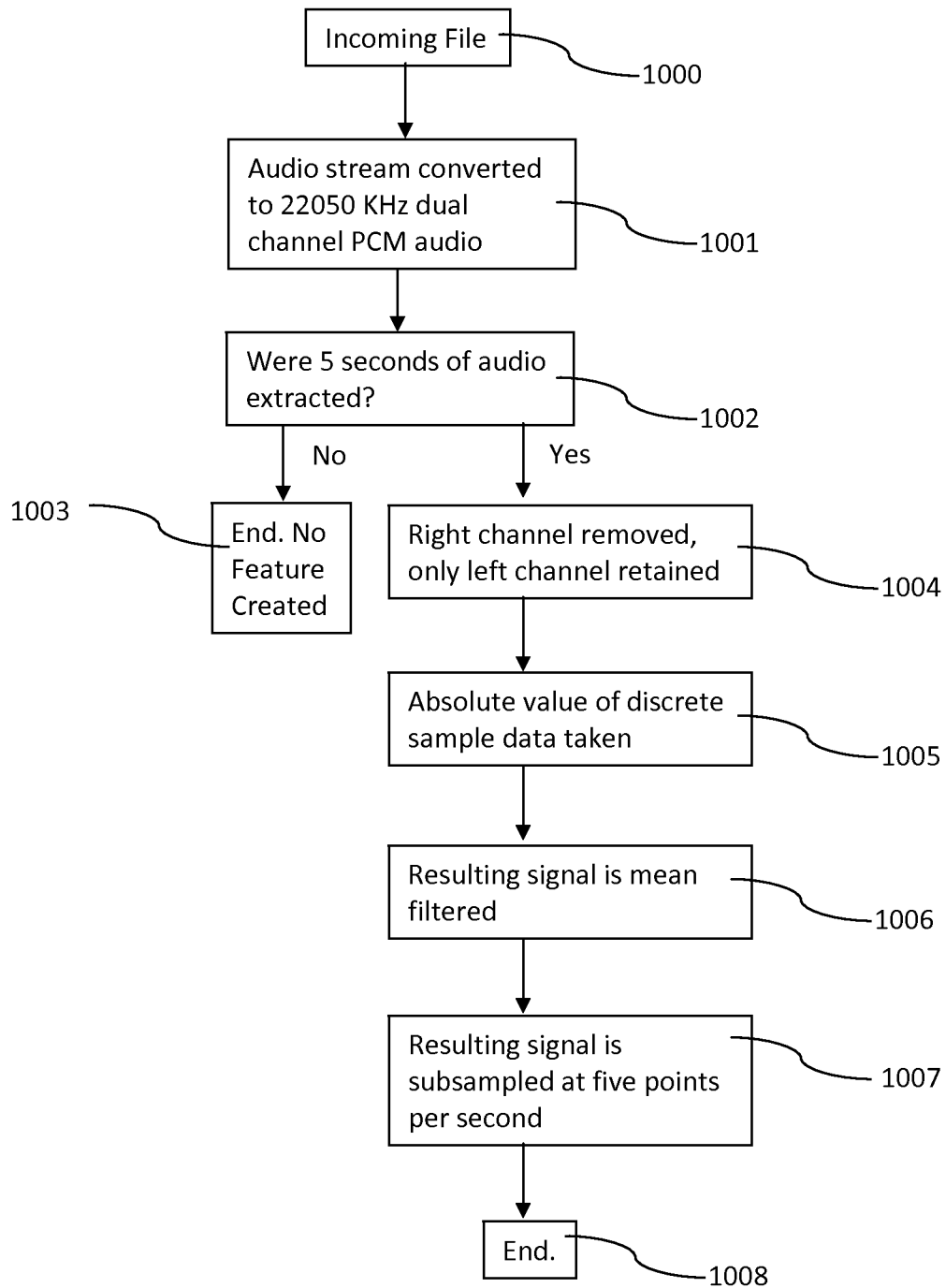
Fig. 8

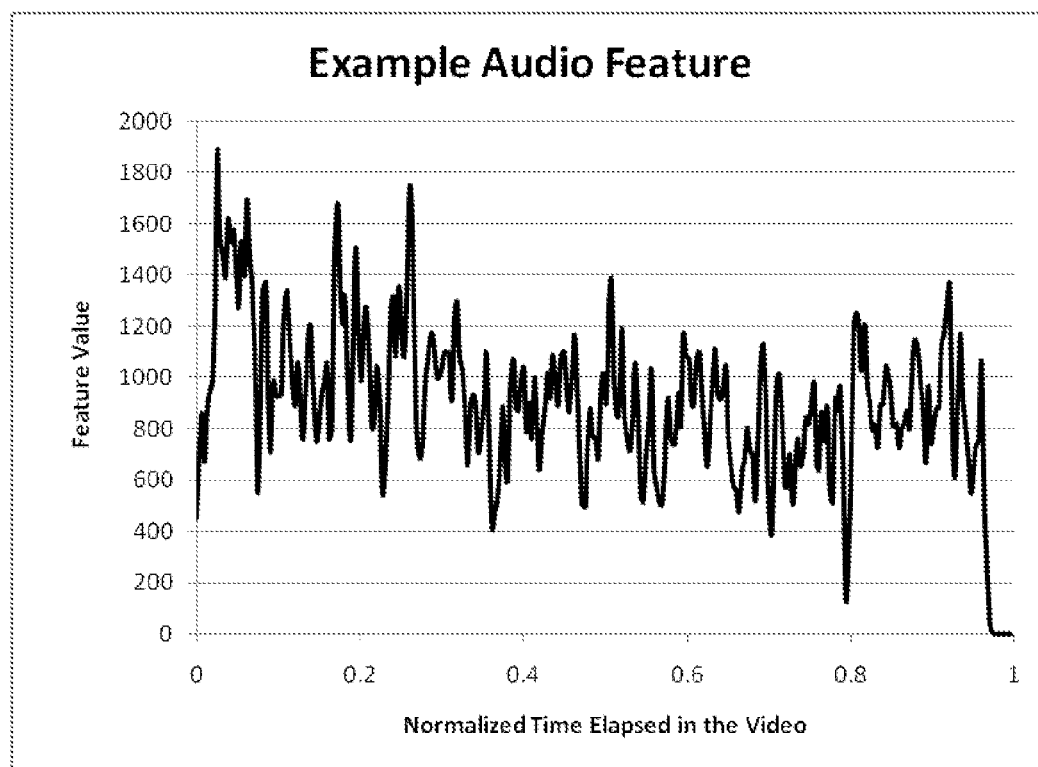


11111111	11111111
11111111	11111111
11111111	11110000
00000000	00000000
11110000	00000000
11111111	11111100
11111111	11111100
11111111	11111100
11111111	11000000
11111111	00000000
11111111	00000000
00000000	00000000
11000000	00000000
11000000	00000000
11111100	00000000
11111111	11111111

255	255
255	255
255	240
0	0
240	0
255	252
255	252
255	252
255	192
255	0
255	0
0	0
192	0
192	0
252	0
255	255

Fig. 9

**Audio Feature Extraction Process Flow Chart****Fig. 10**

**Fig. 11**

## Matching Algorithm Process Flow Chart

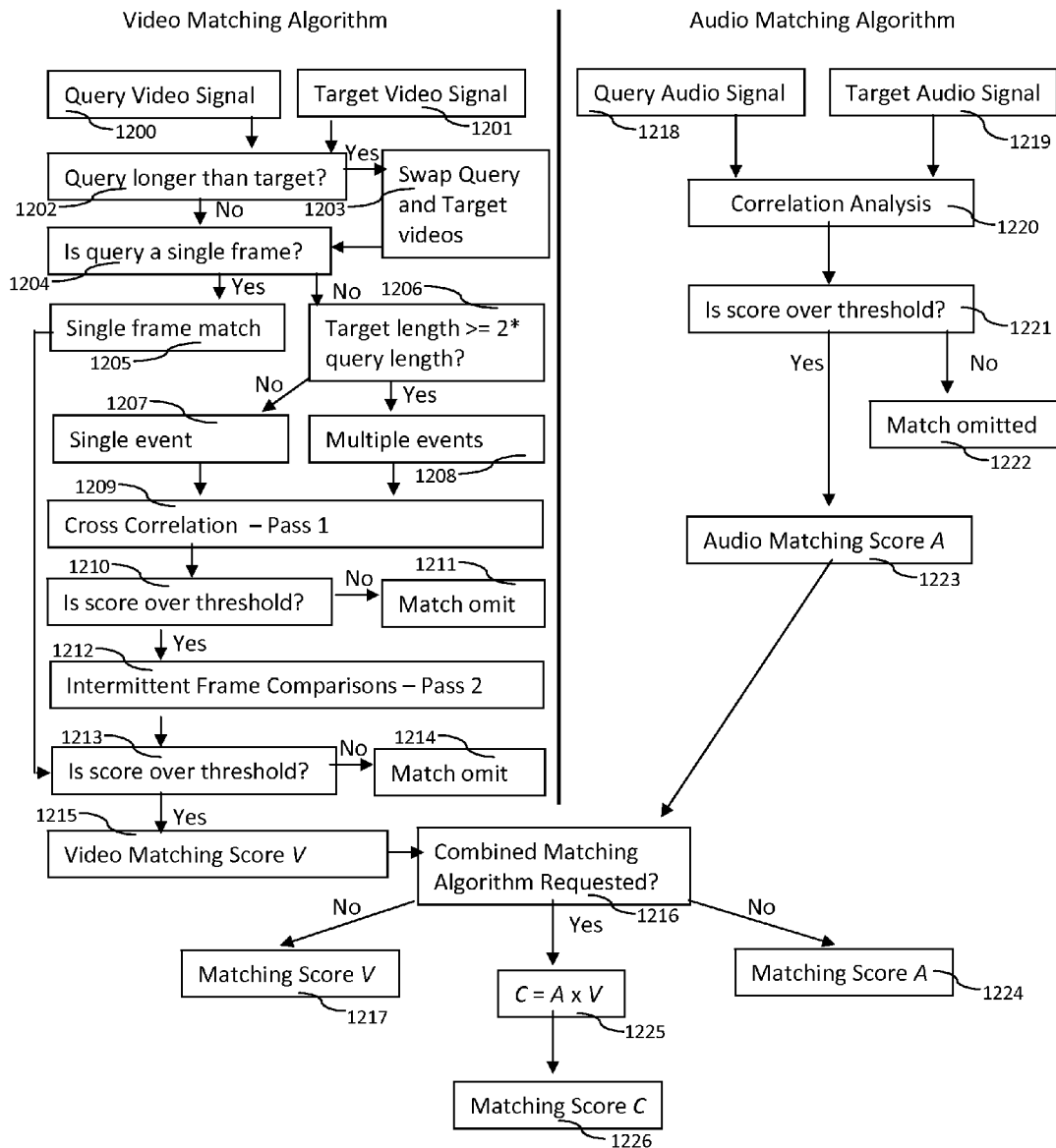


Fig. 12

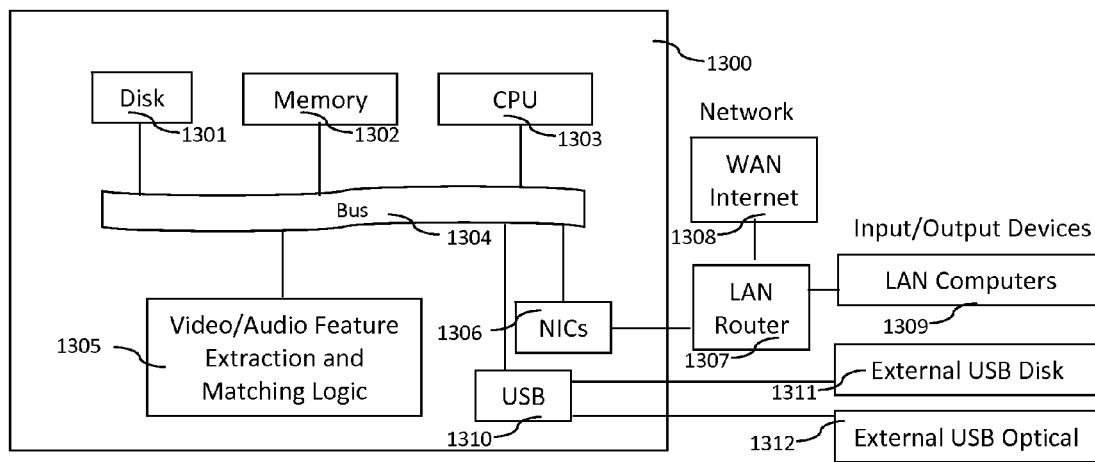


Fig. 13













1) Query MD5	4CCC6F3AEB6C66D76A39E3469891Aa24	Curr Eval File Num 47591
1-1) Video MSec Difference: 0	Clip MSec Difference: 0	Pass 1 Score: 1.000
 Start MSecs: 0	 Clip Duration MSecs: 5000	 End MSecs: 5000
 Start MSecs: 0	 Clip Duration MSecs: 5000	 End MSecs: 5000
1-2) Video MSec Difference: 0	Clip MSec Difference: 0	Pass 1 Score: 0.996
 Start MSecs: 0	 Clip Duration MSecs: 5000	 End MSecs: 5000
 Start MSecs: 0	 Clip Duration MSecs: 5000	 End MSecs: 5000

Fig. 14


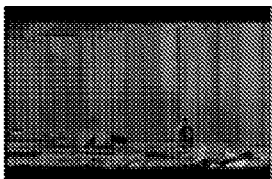


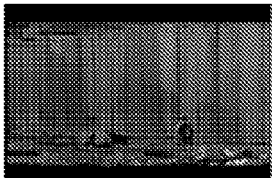
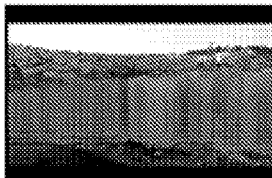


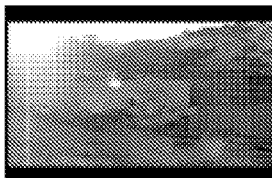


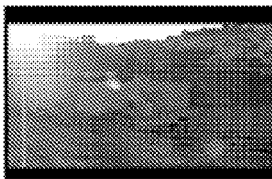
5) Query MP3	151B9F286D92A49811309D3358E06F60	CurEvalFileNum 51546
5-1)Video MSec Difference: 0	Clip MSec Difference: 0	Pass 1 Score: 0.997
 Start MSecs: 0	 Clip Duration MSecs: 310500	 End MSecs: 310500
 Start MSecs: 0	 Clip Duration MSecs: 310500	 End MSecs: 310500
5-2)Video MSec Difference: 140500	Clip MSec Difference: 0	Pass 1 Score: 0.979
 Start MSecs: 0	 Clip Duration MSecs: 170000	 End MSecs: 170000
 Start MSecs: 0	 Clip Duration MSecs: 170000	 End MSecs: 170000

Fig. 15



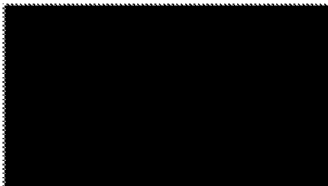
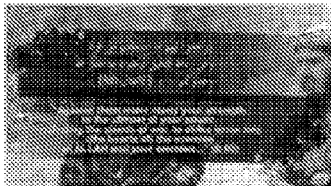

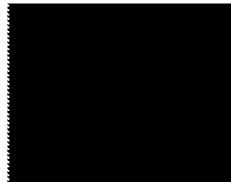
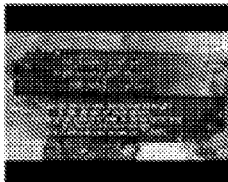
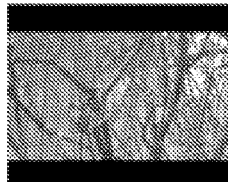
10-9)Video MSec Difference: 1000			Clip MSec Difference: 0		Pass 1 Score: 0.983	
 <p>Start MSecs: 0</p>			 <p>Clip Duration MSecs: 1897000</p>		 <p>End MSecs: 1897000</p>	
 <p>Start MSecs: 0</p>			 <p>Clip Duration MSecs: 1897000</p>		 <p>End MSecs: 1897000</p>	

Fig. 16

## ROBUST VIDEO RETRIEVAL UTILIZING VIDEO DATA

### RELATED APPLICATION(S)

This application is a continuation-in-part of U.S. application Ser. No. 12/538,476, filed Aug. 10, 2009.

The entire teachings of the above application(s) are incorporated herein by reference.

### BACKGROUND

#### 1. Field

The present application relates generally to digital media and more specifically to the process of quickly, efficiently and accurately retrieval similar videos based on extracted feature comparison.

#### 2. Description of the Related Art

Regarding content based video retrieval, one of two main approaches are usually employed. The first is related to matching specific extracted key frames from one video to another. Key frames are extracted at regular intervals, or sometimes selected by scene change detection algorithms. A popular approach is to simply compare key frames of videos using new or existing content-based image retrieval (CBIR). The second is related to modeling the entire clip, and performing a model based comparison during the retrieval. Another existing video retrieval technique is to model entire video clips in some manner, and then perform a model comparison during retrieval. While other models are available, the main model used is a temporal model.

##### Key Frame Comparison

Key frames are often extracted at regular intervals, or sometimes selected by scene change detection algorithms. A popular approach is to simply compare key frames of videos using new or existing content-based image retrieval (CBIR). However, this analysis suffers from two large shortcomings.

Some specific examples of existing technology that utilizes key frame comparison for video retrieval are as follows.

##### 1) U.S. Pat. No. 5,982,979

The video retrieving method provides a video retrieval man-machine interface which visually specifies a desired video out of many stored videos by using previously linked picture data corresponding to the videos. Also, a video reproduction operating man-machine interface visually designates the position of reproduction out of the picture group indicative of the contents. The video retrieving method employs video data, character information linked to the video data, picture data linked to the videos, and time information corresponding to the picture data in the video data. The character information is composed of a title of each video and a creation date thereof. The picture data include, as retrieval information, one picture data representing the content of the relevant video (one picture expressing the video, i.e., a leaflet or the like), and a plurality of picture data adapted to grasp the contents of the entire video. The time information indicates the temporal position of the picture data in the video data.

Hauptmann, A. G., Christel, M. G., and Papernick, N. D., Video Retrieval with Multiple Image Search Strategies, Joint Conference on Digital Libraries (JCDL, '02), Portland, Oreg., pp. 376, Jul. 13-17, 2002 describes the Informedia digital video library which provides automatic analysis of video streams, as well as interactive display and retrieval mechanisms for video data through various multimedia surrogates including titles, storyboards, and skims.

Another existing video retrieval technique is to model entire video clips in some manner, and then perform a model

comparison during retrieval. While other models are available, the main model used is a temporal model.

One example of existing technology that utilizes temporal modeling for video retrievals is in Chen, L. and Stentiford, F. W. M., Video sequence matching based on temporal ordinal measurement, Pattern Recognition Letters, Volume 29, Issue 13, 1 Oct. 2008, Pages 1824-1831. That paper proposes a novel video sequence matching method based on temporal ordinal measurements. Each frame is divided into a grid and corresponding grids along a time series are sorted in an ordinal ranking sequence, which gives a global and local description of temporal variation. A video sequence matching means not only finding which video a query belongs to, but also a precise temporal localization. Robustness and discriminability are two important issues of video sequence matching. A quantitative method is also presented to measure the robustness and discriminability attributes of the matching methods. Experiments are conducted on a BBC open news archive with a comparison of several methods.

Another approach using temporal modeling is described in Chen, L., Chin, K. and Liao, H., An integrated approach to video retrieval, ACM International Conference Proceeding Series Vol. 313, Proceedings of the nineteenth conference on Australasian database—Volume 75, 2008, Pages 49-55. There it is described that the usefulness of a video database depends on whether the video of interest can be easily located. This paper proposes a video retrieval algorithm based on the integration of several visual cues. In contrast to key-frame based representation of shot, the approach analyzes all frames within a shot to construct a compact representation of video shot. In the video matching step, by integrating the color and motion features, a similarity measure is defined to locate the occurrence of similar video clips in the database.

U.S. Pat. No. 7,486,827 describes a two-step matching technique is embodied in a video-copy-detection algorithm that detects copies of video sequences. The two-step matching technique uses ordinal signatures of frame partitions and their differences from partition mean values. The algorithm is not only robust to intensity/color variations it can also effectively handle various format conversions, thereby providing robustness regardless of the video dynamics of the frame shots.

### BRIEF SUMMARY

These prior art approaches each have limitations.

With the key frame approach:

1) If frames are extracted based on a non-temporal basis (i.e., a set number of frames are skipped between each key frame), then differences in frames per second (fps) will cause extracted key frames from similar videos to not align properly, yielding inaccurate results. Furthermore, if the temporal alignment is very close, but not 100% the same, it is possible for a scene change to cause a different frame to be selected for two videos at the same point in time.

2) If a scene change detection algorithm is used to select key frames, the retrieval will only be as good as the scene change detection algorithm, and a propagation of error may be in effect. From experience, it is very rare to see the exact same set of key frames extracted from two videos which were encoded differently from the same source. While the sets of extracted key frames need not be identical, variations ultimately impact relevance ranking in the matched results.

Temporal modeling often suffers from three large shortcomings as follows.

1) Aligning video data in the time domain is not an easy task. Reliably determining an exact frames per second (fps)

value, and then extracting frames uniformly based on that fps becomes a large dimension to the retrieval problem.

2) Videos with little to no motion are very difficult to model in the time domain.

3) Very short videos often yield very little data to be temporally modeled, regardless of content.

To overcome these and other shortfalls, a system and/or method for video comparison can determine if two video signals match by first extracting features from a first and second video signal, cross-correlating the extracted features thereby providing a cross-correlation score at each of a number of time lags; and then outputting an indication of a degree of match between the first and second video signals.

The target video signal is considered to be a candidate match against the query video signal when the similarity score is over the comparison threshold, it may also determine that target video signal matches the query video signal at an interval corresponding to the lag resulting in a highest combined score.

Furthermore, the feature is extracted from each frame the video signals may be based on a mean value of grayscale pixel values in the frame.

In another configuration, the system and/or method may determine if each of the feature-extracted signals is substantially constant over time, and if not, then determining if there is a match from individual frame spatial features extracted from frames, and tested at regular intervals over the match period.

The above discussion of systems and/or methods is meant to be but a brief overview, and not a complete or sufficient description of the present invention, which should be considered as being defined by the claims that follow.

#### BRIEF DESCRIPTION OF THE DRAWINGS

The foregoing will be apparent from the following more particular description of example embodiments of the invention, as illustrated in the accompanying drawings in which like reference characters refer to the same parts throughout the different views. The drawings are not necessarily to scale, emphasis instead being placed upon illustrating embodiments of the present invention.

FIG. 1 illustrates the general process flow and decision process for the video extraction.

FIGS. 2(a) and 2(b) show an example of a letterbox introduced issue. FIG. 2(a) was extracted from the original, 16:9 version of the video stream. FIG. 2(b) was extracted from a version of the video where the original 16:9 stream was re-encoded into a 4:3 format with a severe black border resulting on the top and bottom, and minor black borders resulting on the left and right.

FIG. 3(a) represents the horizontal scan line edge variance and FIG. 3(b) represents the vertical scan line edge variance for FIG. 2(a). FIG. 3(a) and FIG. 3(b) represent the horizontal scan line edge variance and vertical scan line edge variance for FIG. 2(b), respectively. There is no area to remove from FIG. 3(a) and FIG. 3(b), and the area to remove from FIG. 3(c) and FIG. 3(d) is indicated by the shaded, outlined regions 300, 301, 302 and 303.

FIGS. 4(a) and 4(b) show the result of the letterbox cropping for the two frames shown in FIGS. 2(a) and 2(b). FIG. 4(a) was not cropped at all, whereas a significant amount was cropped from FIG. 4(b).

FIG. 5 illustrates an example of one time series signal derived from a video stream.

FIG. 6 illustrates an example flatline signal.

FIG. 7 illustrates the spatial frame edge map feature, and how it is quantized and stored in binary format.

FIG. 8 illustrates the local histogram equalized luminance feature, and how it is quantized and stored in binary format.

FIG. 9 illustrates the global frame color histogram feature, and how it is quantized and stored in binary format.

FIG. 10 illustrates the general process flow and decision process for the audio extraction.

FIG. 11 shows an example of the audio feature signal.

FIG. 12 illustrates the general process flow and decision logic for the video matching, audio matching, and combined video and audio matching algorithm.

FIG. 13 illustrates the video matching technology housed within the compute environment for which it was designed.

FIG. 14-FIG. 16 show select result cases from the technology in operation.

#### DETAILED DESCRIPTION OF THE INVENTION

A description of example embodiments of the invention follows.

##### Introduction

With the increasing presence of online multimedia as well as the wide gamut of formats and encoding schemes available, the need for accurate and efficient video retrieval schemes are of growing importance. Video similarity matching is trivial when both the query and target video generate the same md5 hash value. Additionally, it is not very difficult to compare files in the same video format. In addition, metadata that can be extracted from multimedia files can also make searching a trivial task, as the algorithm becomes an exercise in text comparison. However, when md5 values and metadata either do not exist or are not helpful in determining matching, the only other data to use is derived from the multimedia content itself.

The present system, a Content-based Video Retrieval (CBVR) system, analyzes both video and audio streams from digital multimedia files, and calculates a feature (or features) that uniquely and abstractly describes the media. These features are then compared against each other during the retrieval process in order to establish their "similarity". Unfortunately, several factors can contribute to inaccuracies during the retrieval process, including, but not limited to, changes in file formats and encoding procedures, changes in content quality, and absent video and/or audio data. The possible factors that contribute to retrieval inaccuracy are discussed in more detail below. These factors exist naturally and frequently in real world applications, so it is important to select features that are invariant to these types of irregularities. The present system thus emphasizes tolerance to the variability found in a large generalized corpus of audio and video files.

##### Problem Statement

Given an arbitrary query multimedia file that contains video content, audio content, or both, find other similar multimedia files in a large repository of files in a quick and efficient manner. Similarity, in this context, is defined as multimedia files that may or may not have md5 hash duplicates of each other, do not have matching metadata, but have the same video/audio content differing by one or more of the following characteristics:

The query video or audio data is a subclip of the target video (or vice versa)

The query video and target video differ by quality or integrity (data corruption)

The query video and target video are of differing file and encoding formats

5

The query video and target video contain little to no temporal variation (content does not change much over time, e.g. a stationary surveillance video)  
 The query video and target video differ by frame size  
 The query video and target video differ by aspect ratio  
 The query video and target video differ by color saturation  
 The query video and target video differ by contrast and/or brightness  
 The query video and target video contain no audio information  
 The query asset and target asset contain no video information

#### Current Technical Approach of the Preferred Embodiment

The preferred multimedia retrieval approach utilizes features from both digital video and audio data. In a specific example, digital video, as well as digital audio (such as in the form of extracted dual channel 22050 Khz Pulse Code Modulated (PCM) data) are used. They are first correctly aligned in the time domain. Features are then extracted from the data and temporal signatures are created.

In addition to a temporal signature, specific spatial features are computed for each frame. A combination of cross correlation analysis (in the case of temporal features) and direct bit-wise comparison (in the case of spatial frame features) are then used during the retrieval process to match extracted features of digital media to one another.

#### Feature Extraction

For each media asset, an attempt to extract both audio and video features takes place. If either the audio or video stream is unavailable or encoded with an unrecognized codec, no feature extraction is possible for that stream. In addition, if the extracted stream yields 0 frames, no feature extraction is attempted for that stream. In other words, at least 1 frame must be extracted from the video for visual feature extraction to take place.

#### Video Feature Extraction

FIG. 1 illustrates the general process flow and decision process for the video extraction. A very brief summary of the process is as follows: In step 100, a new, incoming video file is presented to the system. In step 101, an attempt is made to extract frames from the video and a decision is made depending on the outcome. If at least 1 frame was not able to be extracted, in step 102 the process is terminated with no video feature created. Otherwise, step 103 performs a letterbox cropping filter on each extracted frame. Step 104 forces each frame to be resized to 318x240. Step 105 extracts a single, statistical measurement from each frame. Step 106 computes 3 separate spatial features for each frame. In step 107, the single frame measurement is converted into temporal signals. In step 108, a flatline analysis is performed on the global frame temporal signal. In step 109, the video featuring is completely done.

The following is a more detailed summary of the step-by-step process for video feature extraction:

- 1) In step 101, JPG frames are extracted from the original media file at a rate of 1 frame every 500 ms (or 2 frames per second). If at least one frame cannot be extracted, the process is aborted and no feature is created.
- 2) For each frame extracted during step 101, a letterbox cropping filter is applied (step 103) to the frame to remove a possible border artifact. This is performed on frames that contain a black border and will increase the likelihood of matching the frame with a black border to the original video because both frames have different

6

aspect ratios. FIG. 2 shows an example of a letterbox introduced issue. FIG. 2a was extracted from the original, 16:9 version of the video stream. FIG. 2b was extracted from a version of the video where the original 16:9 stream was re-encoded into a 4:3 format with a severe black border resulting on the top and bottom, and minor black borders resulting on the left and right.

The Letterbox cropping filter first performs an edge analysis on the frame. There are many edge detection algorithms available, for example, the Sobel edge operator described in Sobel, I., Feldman, G., "A 3x3 Isotropic Gradient Operator for Image Processing", presented at a talk at the Stanford Artificial Project in 1968, unpublished but often cited, orig. in Pattern Classification and Scene Analysis, Duda, R. and Hart, P., John Wiley and Sons, '73, pp 271-2, and the Canny edge detector described in Canny, J., A Computational Approach To Edge Detection, IEEE Trans. Pattern Analysis and Machine Intelligence, 8:679-714, 1986.

However, for speed and efficiency, a simple edge detection algorithm based on local pixel variance over a threshold can be used. With this approach, a horizontal and vertical scan line variance analysis of the resulting edge map is conducted. Continuous regions of low scan line variance measured from the edge of the frame are discounted from feature consideration. A scan line standard deviation of 0.05 is used as a threshold. This value performs well at detecting actual edge in video frames, while suppressing possible minor false edges that may be an artifact of JPG compression.

In FIG. 3, panels a and b represent the horizontal scan line edge variance (panel a) and vertical scan line edge variance (panel b) for FIG. 2a. Panels c and d in FIG. 3 represent the horizontal scan line edge variance (panel c) and vertical scan line edge variance (panel d) for FIG. 2b. There is no area to remove from FIGS. 3a and 3b, and the area to remove from FIGS. 3c and 3d is indicated by the shaded, outlined regions 300, 301, 302 and 303. FIG. 4 shows the result of the letterbox cropping for the two frames shown in FIG. 2. FIG. 4a was not cropped at all, whereas a significant amount was cropped from FIG. 4b.

Because the intended use of the filter was to remove possible borders, some rules have been incorporated to prevent too much cropping from inadvertently happening. The first rule is that if the total left and right combined removal amount is more than 35% of the total frame width, the frame is untouched. Similarly, the second rule is that if the total top and bottom combined removal amount is more than 35% of the total frame height, the frame is untouched. These amounts were derived by taking each expected worst possible case, a 16:9 forced into a 4:3 resulting in black top/bottom borders (rule 2), and a 4:3 forced into a 16:9 resulting in black left/right borders (rule 1), plus a little bit extra for expected error.

- 3) In step 104, the resulting letterbox filtered image then undergoes a forced resize to 318x240 for two reasons: 1)

The computed features are very spatially generic, and do not suffer from reduced quality or lack of fidelity at the local pixel level, and 2) Reducing large frame sizes (sometimes as high as 1000x500 and larger) greatly reduces feature extraction and further processing time.

- 4) From step 104, once each frame has been resized, then the actual feature extraction takes place. The global frame feature in step 105 is a single point measurement that uniquely describes the frame that we are analyzing, so a measurement was picked that incorporated global pixel averages and global pixel standard deviations.

Mean value of frame pixel grayscale values divided by 255.0  
 Stdev value of frame pixel grayscale values divided by 127.5

The denominators in each calculation normalize each value to [0.0, 1.0]. The goal of this feature extraction measurement is to represent the distribution of target pixel values as simply and as descriptively as possible by a single, scalar value. As statistical distributions are described by their moments, the two components (stdev and mean, respectively) represent the first two moments of the distribution. However, while both features are being extracted (mean, stdev), currently only the mean value is being used during matching, as this supplies sufficient discrimination for an initial pass.

5) At step 107, when the preceding has been done for all frames extracted for the video, this yields 2 time series signals (time vs. measurement). In all, there are two signals derived from one video stream, the signal for the 2 global frame measurements. FIG. 5 illustrates an example of one time series signal derived from a video stream.

6) As will be described below, the retrieval process relies on differences and uniqueness in these derived features. However, as is sometimes the case, video streams do not have a lot of signal change, especially short videos. For example, a 15 second clip captured from a surveillance camera showing little to no activity (as detected in step 108) would have little to no change in frame to frame point measurements, thus resulting in a "flatline" signal. See FIG. 6 for an

A signal is identified as being a flatline signal based on both the statistical distribution of the temporal data points as well as the length of the video. The longer the video and the more varied the temporal point data is, the greater the chance that a video will not be tagged as a flatline.

Component 1=mean value of calculated global frame means (temporal feature 1)

Component 2=mean value of calculated global frame stdevs (temporal feature 2)

Specifically, a video is tagged as flatline when the following condition is met:

$$1.0 > \{(\text{numFrames} * \text{Component 1}) + (\text{numFrames} * \text{Component 2})\} / 2.0$$

Flatline analyses on video streams are used and are useful in aiding logic and branching path decision during the matching process.

In addition to a single point measurement being calculated on the global frame, a set of 3 features are being extracted from each frame: localized detected edge, localized luminance-based feature, and a global color histogram in step 106 on the frame. The specific spatial frame features are described in detail below:

#### Localized Edge Detection:

Using the same edge detection algorithm used in the letterbox detection technique outlined above, edge is detected on the frame and an edge map of the frame is created. The edge map is quantized into a 16x16 grid, and the edge map grid is then encoded into a binary matrix of 0's and 1's, and stored numerically until comparison. FIG. 7 shows an example of the edge map feature, and how the edge map is quantized into a grid, and each cell is encoded into a binary map and stored numerically

#### Localized Luminance:

The frame is converted to grayscale, and then a histogram equalization is computed to correct for variations in luminance and contrast. The histogram equalized grey values are then quantized into a 16x16 grid, and the grid is encoded in binary such that the upper 50% of values are encoded as a 1, and the bottom 50% of values are encoded as a 0. The matrix is then stored numerically until comparison. FIG. 8 shows an

example of the histogram equalized luminance feature, and how the luminance is quantized into a grid, and each cell is encoded into a binary map and stored numerically.

#### Global Color Histogram

A 16-color global color histogram is computed on each frame. The frames colors are quantized into 16 bins based on closest proximity in 3D RGB space, a histogram is created based on the frequency of each of the 16 quantized colors, and finally this histogram is represented in binary, and is stored numerically until comparison. FIG. 9 shows an example of the global color histogram feature, and how the color is encoded into a binary map and stored numerically.

#### "Blank" Frames:

An issue with these frame features, how they are computed, and how they are compared to each other is the notion that blank or near blank frames (i.e., a blank frame with a single white word across the center), actually compare very, very highly, even though the actual content of the frame is completely dissimilar. We have developed a concept of a "blank" frame, or more specifically, these 3 frame spatial features are only computed if there is deemed enough valid frame content to perform a meaningful frame comparison. A frame is determined to be "nonblank" if the global standard deviation of its grayscale pixels is >25, and the amount of pixels that contain local pixel variance is 35% of the total image area. These values were determined based on a set of training images (blank, nonblank) that were fed to a trivial linear classifier. This aids in the problem of blank or near blank frames returning high matching scores, even though they have no relevant content.

#### Audio Feature Extraction

FIG. 10 illustrates the general process flow and decision process for the audio extraction. A very brief summary of the process is as follows: In step 1000, a new, incoming video file is presented to the system. In step 1001, the audio stream of the incoming file is extracted as a 22050 KHz dual channel PCM audio file. In step 1002, a decision is made depending on the outcome of step 1001. If 5 seconds worth of audio data was not able to be extracted, in step 1003 the process is terminated with no audio feature extracted. Otherwise, in step 1004 only the left channel is retained for analysis. In step 1005 the absolute value of the raw audio data is taken. In step 1006 the audio data is filtered with a half-second running mean filter. In step 1007, the data is subsampled at 5 points a second. In step 1008, the featuring is complete and the feature is built.

The following is a more detailed summary of the step-by-step process for audio feature extraction:

1) In step 1001, a 22050 KHz PCM (WAV) dual channel audio file is extracted from the original media file. If no WAV file is created, or less than 5 seconds of WAV data is created, the audio feature generation is aborted and no audio feature is created in steps 1002 and 1003.

2) In step 1004, only the left channel is retained for the feature extraction, the right channel is disregarded. For this instantiation of the system, we only use the left channel to avoid issues with having to mix the left and right channels together to form one audio signal prior to feature extraction. For another instantiation, the right channel could be used.

3) At step 1005, only the absolute value of the sample data is considered. Audio signals, by nature, have amplitude values that oscillate inversely on either side of 0. Taking the absolute value allows the feature space to be completely positive, while preserving the actual magnitude of the amplitude,

- 4) A running window mean filter is applied in step **1006** to the left channel data with a size of 11025. At 22050 Khz, this window corresponds to a half second. A half second window was chosen as a good benchmark for the minimum average length of silence between spoken phrases.

This allows brief periods of silence in speech to have minimal amplitude, even when mean filtered. Mean filtering also allows audio signals to correlate well together, even if the source audio streams differed in dynamics or quality.

- 5) At step **1007**, the resulting signal is approximated by subsampling the data, 5 samples per second. This subsampling greatly improves the speed of the comparison, while still providing good signal structure, uniqueness, and discriminatory properties. 5 samples per second were chosen for this instantiation although other sampling rates could have been chosen.
- 6) The final audio feature is written in binary format. As opposed to the video features, the audio feature values are not normalized to [0.0, 1.0]. FIG. **11** shows an example of the audio feature signal.

#### Feature File System Representation

To optimize the system's processing and retrieval speed, the extracted features are created and stored in binary flat files for quick distribution, access and maintenance

Video and audio feature data is represented in one of two ways during the ingestion process. The first manner is referred to as a signature file, which is a binary encoding for a specific feature for one specific video or audio stream. The second manner is referred to as a repository file, which is a concatenated collection of signature files, with some additional header information that describes the size of the repository.

#### Signature Files

Each feature extracted, either video or audio, is written in a binary file format. These "signature" files contain some metadata regarding the file and the feature, as well as the feature itself. The signature files contain a 58 byte header which contains metadata regarding the extracted feature, followed by a variable amount of actual feature data. The header contains a 32-byte string that contains the complete md5 hash value for the media file. This is the primary lookup index for this file. Also included in the header are 2 flags, an active flag to dictate whether or not the signature is active, and a flatline flag to identify certain signatures as flatline (see elsewhere for more information on flatline features). The last 3 values in the signature file are values for the number of features in the signature file, the number of data points in the data block, and the number of samples per second for the data block. The last

allows for a temporal mapping of data to time. Following the header is the actual data block. As described elsewhere, all features contain a dynamic amount of feature data that is a function of media length. Table 1 illustrates the format of the feature signature files.

TABLE 1

Description of a binary feature signature file						
Signature File Header						Signature
Bytes 1-32	Byte 33	Byte 34	Bytes 35-42	Bytes 43-50	Bytes 51-58	File Data Bytes 59-
Complete File MD5	char value: Active 0/1	char value: Flatline 0/1	Unsigned 64 bit int value: Number of features	Unsigned 64 bit int value: Number of points (NP)	Double float value: Data points per second	Double float or unsigned char values: Feature Data

#### Repository Files

When a new audio or video file is introduced to a content management system utilizing the techniques described herein, individual signature files are created, and then appended to existing repositories. A repository file has a header that describes the size of the repository, and this header is updated every time a signature is added to or removed. In addition, to better manage the size of repository files, as well as aid in the parallel distribution and accessing of very large repositories, a separate repository is created for each possible hex value (0-F), and signatures with md5 values that begin with a particular hex are only inserted into the repository matching that hex value (via repository file name). Table 2 illustrates the format of the repository files.

TABLE 2

Description of a binary feature repository file	
Header, 8 byte unsigned 64 bit int value: Number of signatures	
Signature File 1	
Signature File 2	
Signature File 3	
...	
Signature File N	

Because the header information for each signature file is intact in its corresponding repository file, it's possible to traverse individual signatures in a repository by calculating byte offset information from the signature header, and then moving to the next offset where the next signature starts. However, because signatures can be of varying data lengths, repository serial access is possible, but random access is impossible unless you know the direct offsets to seek to ahead of time.

When signatures are removed from a repository file, instead of physically removing the signature from the file and completely rewriting the end of the repository file, the Active flag found in the head of the signature to be deleted is set from 1 to 0. Signatures flagged as 0 can be quickly skipped when the complete repository is traversed during the retrieval process.

#### Matching Algorithm

FIG. **12** illustrates the general process flow and decision logic for the video matching, audio matching, and combined video and audio matching algorithm. A brief description of the process is as follows:

## 11

For video matching, query video features **1200** and target video features **1201** are loaded. In step **1202**, the question is asked if the query length is longer than target, and if so, the query and target are swapped (**1203**) for the current comparison. This is performed to simplify case implementation, as the shorter video is always compared to the longer video, regardless of which was the query and which was the target. Next, if the query is a single frame (**1204**), then in step **1205** a special retrieval case is invoked for single frame queries. Since cross correlations of single points are not statistically valid, the spatial frame features are then compared. If the query is not a single frame, then in step **1206** the question is asked whether or not the query length is less than 2 times the target length. If it is not, then a single matching event is searched for (step **1207**). Otherwise, possible multiple matching events are possible (step **1208**). In step **1209**, a cross correlation is performed on mean temporal feature (pass 1). If any event(s) have a matching score greater than or equal to the pass 1 threshold (**1210**), then the possible matching event(s) move onto a second pass (**1212**). Otherwise, the events are omitted (**1211**). Step **1212** consists of intermittent frame comparisons, to validate the matches at a spatial frame level, and for the entire duration of the match. If the intermittent frame score is greater than the pass 2 threshold (**1213**), then the match is retained as a visual match with score V (**1215**). Otherwise, the match is omitted (**1214**).

For audio matching, query audio features **1218** and target audio features **1219** are loaded. In step **1220**, matching is performed on the features. If they are not over the matching threshold in step **1221**, then the target is omitted as match in step **1222**. Otherwise, the match is retained in step **1223** with audio matching score A. If the combined matching score is not requested in step **1216**, then either the video matching score **1215** or the audio matching score **1224** is the final score. Otherwise, the product of A and V is computed in step **1225** and represents the final combined matching score C in step **1226**.

A more detailed description of the matching algorithm follows.

The baseline algorithm for matching any temporal feature signatures is a cross correlation analysis. The cross correlation analysis is a standard digital signal processing technique for finding signals embedded within other signals with a high degree of correlation at any lag delay. For example, if one signal was 10 data points in length, and another 50 points in length, the cross correlation would test the correlation at every possible matching point (in this case, the smaller signal could match the longer signal at any one of 40 matching points). The general equation for the cross correlation of two arbitrary 2D signals is:

$$(f \star g)[n] \stackrel{\text{def}}{=} \sum_{m=-\infty}^{\infty} f^*[m]g[n+m].$$

The cross correlation is very good magnitude independent way of comparing signals to one another. However, one drawback to the use of the cross correlation analysis in this application is that very short videos yield very short signals (as few as 10 data points), and it's possible for a very short signal to randomly correlate very highly to a long signal, even if the original videos that the signals were extracted from contain no similar content. Thus, the individual spatial frame features are used to help in these cases, as a secondary pass.

The spatial frame features are binary in format, and encodings of spatial information relevant to characteristics of the

## 12

frame. The features are stored as collections of binary unsigned char values, aligned with each other. The comparison is a loop through all numbers relevant to both frame spatial features, an exclusive OR operation on each number pair, and a counting of the number of '1' bits in the result. This total, then, is divided by the total number of bit locations possible. For example, there are 768 total bits in each frame comparison. If the total number of '1' bits after all of the exclusive OR operations was 47, then the frames are said to be 721/768 relevant, or 93.8% relevant.

#### Video Matching Algorithm (steps **1200-1218**)

The algorithm to match video features is based on the preceding matching algorithm. There are several independent branches of matching logic that are invoked, depending on properties of the query and target video segments. Each branch will be described in detail below:

##### Single Frame Query Match:

Sometimes there are cases when a single query video frame is submitted for search. Reasons for this would include: 1) That's all the user supplied for the query, and 2) only one frame could be extracted from the video. In these cases, the temporal feature will only contain one point, and the cross correlation of single points is statistically meaningless. Thus, these types of searches go straight to the spatial frame features. For each frame in the target video, the frame is compared to the query frame, and contiguous ranges of target frames that are above the pass 2 threshold are reported as single events in the output matching results. The score for the event is computed as the average of the frame comparison matches that comprised the event.

##### Lengthy, Non-Flat Line Searches:

There is sufficient reason to believe that for lengthy videos whose temporal signals contain enough variability and unique "structure" (i.e., non flat line), that the cross correlation by itself is sufficient to determine accurate matches and minimize false positives. The length to determine this condition is parameterizable, but though empirical trials were determined to be 5 minutes (300 frames). More specifically stated, if the lengths of both the query and target video segments are both  $\geq 5$  minutes, and both temporal signals are non-flatlines, then the cross correlation alone is sufficient to determine an accurate result set. The spatial frame features are not needed.

##### Single Event Matching:

If the query video length is greater than half the target video length (or vice versa), then it can logically be deduced that one and only one matching event should be searched for. In this case, a cross correlation on the temporal signals is conducted, and if an event passes the pass 1 threshold, then a secondary pass (pass 2) involving an intermittent spatial frame comparison is conducted. At 20 regular intervals during the match (every 5% of the match duration), a spatial frame comparison is conducted. Frames that compare with scores higher than the pass 2 threshold keep the match alive. Conversely, frames that compare with scores lower than the pass 2 threshold have the capability of permanently cancel the match. Along with pass 2 is the notion of "frame misses", and "frame miss tolerance". Frame misses are the total number of continuous frame comparisons that missed the pass 2 threshold. Frame miss tolerance is the number of continuous frame misses to tolerate before the match is cancelled. A frame that compares higher than the pass 2 threshold will reset the number of current frame misses back to zero. If the intermittent frame comparisons reach the conclusion of the match duration without violating the frame miss tolerance, then the match is retained permanently with a score that is the average of the frame comparisons that were not "misses"

## 13

## Multiple Event Matching:

This mode of search is very similar to the single event matching described above. Basically, one or more instances of the single search method are used on completely disjoint sections of the query/target pair. Once a positive match is found, that section is removed from further consideration, and other areas are scanned for possible matches, until the entire sequence has been exhausted.

## Match Output Format:

In all cases, the match output format is the same, regardless of which matching logic branch produced the result:

The query md5 value

The start and end time that the query video matched the target video

The target md5 value

The start and end time that the target video matched the query video

The final matching similarity score

Audio Matching Algorithm (steps 1218-1224)

The audio matching algorithm is simply the general matching algorithm described above for one query signature against a repository of signatures, given an input matching similarity threshold. However, for this match, only the correlation score is used as the matching score. Because the audio features are not normalized, the mean magnitude difference score is not applicable. The output is a list exactly the same as the list produced for pass 1 of the video matching algorithm. However, in the case of audio, there is only one pass. The output list of matches is final.

Combined Video and Audio Matching (Step 1216 and Those Following)

In cases where the query file has both audio and video data, and thus both audio and video signatures, it is possible (and recommended) to couple the audio and video retrieval matching scores together, to further discriminate similar videos from non similar videos. The method to combine the audio matching score A (step 1224) with the video matching score V (step 1215) into a combined score C (step 1226) is defined as:

$$C=A \times V$$

Several combination algorithms were examined. The first, which was  $C=\max(A, V)$ , performed well at honing in on matching videos that had either high matching video scores or high matching audio scores. However, it suffered at accurately ranking, for example, matching videos that had similar video quality but differing audio quality. A video with a high matching video score and high matching audio score should have a higher combined matching score than a video with a high matching video score and a lower matching audio score.  $C=\max(A, V)$  did not allow for this.

The second combination algorithm examined, which was  $C=\text{avg}(A, V)$ , performed well at providing a general picture at how the video matched on a combined audio and video level. However, the variance between A and V is masked. For example, if a video had a matching video score of 0.99 and a matching audio score of 0.91, using  $C=\text{avg}(A, V)$ , this is still an admirable combined matching score of 0.95. However, this masks the fact that the video score matched as low as 0.91, which is not ideal.

$C=A \times V$  was settled on, ultimately, because it scales the disparity between similar and non-similar videos geometrically, not linearly. More specifically, similarity scores in audio and video get amplified when multiplied together, further pushing similar results away from dissimilar results. For example, if one video matched with a 0.98 matching audio score and a 0.96 matching video score, using  $C=A \times V$ , the

## 14

combined score would still be a high 0.94. However, if another video matched with a 0.95 matching audio score and a 0.92 matching video score, using  $C=A \times V$ , all of a sudden the combined score is dropped to 0.87.

## Retrieval Process

Four modes of searching capability are offered:

- 1) Video only search—only the video features of target files are searched against
- 2) Audio only search—only the audio features of target files are searched against
- 3) Audio or video search—On a target file by target file basis, if both the audio and video features are present, the retrieval uses a combined audio/video search for the asset. If not, then the search uses either the audio or the video, whichever is available.
- 4) Audio and video search—both the audio and video features of target files are searched against, and targets that do not specifically contain both are excluded from consideration.

FIG. 13 illustrates the video matching technology housed within the compute environment for which it was designed. System disk 1301, memory 1302, CPU 1303 and the video feature extraction and matching logic 1305 are all interconnected within the data processing environment 1300 by a bus 1304. It should be understood that other elements of a system may be present, such as input devices that provide video/audio data sources and output devices such as displays and audio speakers. However these elements are not critical to the operation of the matching techniques described herein. The data processing environment communicates to local external devices via a Universal Serial Bus (USB) port 1310, and communicates with other networked devices via one or more Network Interface Cards (NIC) 1306. The NIC(s) are connected directly to a Local Area Network (LAN) router 1307. In turn, the LAN router allows connectivity to other LAN computers 1309 as well as the outside internet 1308. The input source videos for the retrieval technique would originate from internet downloads via HTTP or FTP 1308, other networked LAN computer disks 1309, local USB external disk 1311 or local USB optical disk 1312, either via CD or DVD. Output retrieval results might be stored on local disk 1301, stored locally on external USB devices 1311 1312, stored on network computer disk 1309, or served in real time to internet-based clients 1308.

## Results

Three retrieval cases are presented below, illustrating different challenges that the technique attempts to solve.

Case 1: Query video is a short video with no audio stream and very little video content variability. FIG. 14 is a screenshot of a retrieval application based on the technique described in this document after a search was conducted using a short, 5 second video with very little frame-to-frame change from a stationary camera. The search brought back all 3 other versions of the video, differing not only by file type and quality, but also frame resolution. In addition, no false positives were brought back as matched in this result set.

Case 2: Query video is a longer video with audio. FIG. 15 is a screenshot of the same retrieval application based on the technique described in this document. The search brought back 3 other versions of the video, differing not only by file type and quality, but it also brought back a version of the query video that was 50% the duration. This particular matching result (5-6) was a subclip of the original query video. In addition, no false positives were brought back as matched in this result set.

Case 3: Query video is a longer video with audio. FIG. 16 shows a screenshot of the same retrieval application based on



15

the technique described in this document. The search brought back all 11 other versions of the video, differing not only by file type and quality, but also by letterbox and aspect ratio encoding. In addition, no false positives were brought back as matched in this result set.

Even though the size of these example corpuses is small (much, much smaller than the real-world corpuses for which the technique was designed), the nature of the features being computed and the discriminatory nature of the matching algorithm will not falter to added noise from a larger sized repository.

### CONCLUSION

Existing techniques utilized in academia and industry attempt to solve the problem of video retrieval. However, many of them are based on either 1) key frame selection and

16

in retrieving corrupt versions of videos given non corrupt versions, or vice versa. The proposed technique is also robust at accurately detecting videos containing little to no motion, regardless of video length. This is possible by use of spatial feature, which is used when two videos have been identified as having little or no content variation (flatline). The technique is multi-dimensional, meaning that it is robust at using video features when only video features are available, and likewise for audio features, but will utilize both audio and video features, when present, for higher accuracy. The technique is also robust at finding subclips within video files, as the cross correlation analysis attempts to find smaller signals embedded within longer signals. Lastly, the technique is invariant to letterbox-encoded issues, due to the specific letterbox cropping filter employed See Table 3 for specifics.

TABLE 3

Listing of specific searches, and what type of challenge they overcome.			
Specific Challenge Area	Query Number	Query Video MD5	Matched Video MD5
Different formats	7	2399288E6B2292350F20DFD10C5AADCBC	C799D4A31FE896825DC020537D35A1D0
Good video to a corrupt video	4	1E7496F3A18A8F941209A41605039CA6	6E5AA1131021246980C2E71A9481EC0C
Little to no motion	21	BAF2E6AEF34ECFB3C7F0FA4C04B04395	FE9658D977FAA073D4DBD179AA828456
Durations may be slightly off	69	B40286E66A9EF70C3BF1B1A62B7634CB	BDD52B33E46357EDDC5A8F059D4006C5
Containing only video	Many		
Containing only audio	Many		
Clip found within a much larger video	75-1	660AD92C044E5807BC05944CEC50A6AC	8F9E7735BE6B4AE0739BEC4A40DC6C5A
Changes in aspect ratio or letterbox	Many		

comparison, which can suffer greatly from inaccuracies in the key frame selection process as well as forcing a domain-specific solution (image retrieval) on another domain (video data), which is not scientifically optimal, and 2) a temporal representation of the video, which can suffer when the content of the video does not yield a unique temporal signature. The preliminary results and evaluation indicate that the PFI technique proposed here is robust in retrieving similar videos and is invariant to many of the real world challenges described in the problem statement earlier in this document, and is therefore novel and superior to many of the existing video retrieval techniques currently employed today.

Specifically, the proposed technique is invariant to retrieving videos of different formats, as the nature of the video feature, specifically the single point scalar measurement per frame or grid region, are robust to differences in frame encoding and quality. This also allows the technique to be invariant

What is claimed is:

1. A method for determining if two video signals match, the method comprising:
  - extracting features from a first video signal to provide a first feature-extracted signal;
  - extracting features from a second video signal to provide a second feature-extracted signal;
  - performing a first pass of determining if the first feature extracted signal matches the second feature-extracted signal based on a cross-correlation of a temporal feature of the first and second feature-extracted signals at each of a plurality of time lags, to provide a temporal feature match score;
  - determining if each of the first feature-extracted signal and the second feature-extracted signal is a flat line signal as determined from a statistical attribute and a length of the respective feature-extracted signal;
  - if the first feature-extracted signal and the second feature extracted signal are not a flat line signal, then
    - determining a similarity score from only the temporal feature match score without determining a spatial

17

match score between the first feature-extracted signal and the second feature-extracted signal;

else

if the first feature-extracted signal and the second feature-extracted signal are a flat line signal,

then

performing a second pass of determining if the first video signal matches the second video signal based on comparison of a spatial features extracted from both the first video signal and from the second video signal; and

determining a similarity score from a result of the spatial feature match score.

2. The method of claim 1 additionally comprising:

resizing the first and second video signals to a common frame size prior to the feature extractor steps.

3. The method of claim 1 further comprising:

providing an indication of a degree of match between the first and second signals.

4. The method of claim 1 wherein determining a similarity score further comprises:

linearly combining the cross-correlation score at each lag to provide a combined score for each lag; and

determining a highest combined score across a plurality of lags.

5. The method of claim 4 wherein linearly combining comprises weighting the cross-correlation scores equally.

6. The method of claim 1 wherein

the first feature-extracted signal is derived from a feature extracted from each frame of a query video signal and the second feature-extracted signal is derived from a feature extracted from each frame of a target video signal among a plurality of target video signals;

and the method further comprises:

determining if the target video signal is a candidate match for the query video signal by comparing the similarity score to a threshold; and

if the target signal is a candidate match, determining if the target video signal matches the query video signal at an interval corresponding to the lag resulting in the highest combined score.

7. The method of claim 6 wherein the feature is extracted from each frame of the query video signal and of the target video signal based on a standard deviation of grayscale pixel values in the frame, a mean value of grayscale pixel values in the frame, and a difference between mean values of respective red and blue pixels in the frame.

8. The method of claim 7 wherein the feature is computed as a normalized sum.

9. The method of claim 1 wherein the second video signal is confirmed as a match for the first video signal only if all of the cell features attain respective maximal values substantially simultaneously.

10. The method of claim 1 additionally wherein

the first and second video signals are audio-video signals, each audio-video signal comprising audio and video data;

the first and second feature-extracted signals are derived from a video feature extracted from respective first and second audio-video signals; and

the similarity score is a video similarity score;

and the method further comprises:

cross-correlating third and fourth feature-extracted signals derived from audio features extracted from respective first and second audio-video signals to provide an audio cross-correlation score at each of a plurality of lags;

18

determining an audio similarity score between the third and fourth feature-extracted signals based on the audio cross-correlation scores;

multiplying the video similarity score and the audio similarity score to provide an audio-video similarity score; and

determining if the first audio-video signal matches the second audio-video signal by comparing the audio-video similarity score to a threshold.

11. The method of claim 1 wherein the first feature extracted signal is a signal  $f[n]$ , the second feature-extracted signal is a signal  $g[n]$ , a lag amount is given by  $m$ , and the step of cross-correlating comprises determining:

$$(f \star g)[n] \stackrel{\text{def}}{=} \sum_{m=-\infty}^{\infty} f^*[m]g[n+m].$$

12. The method of claim 1 additionally comprising:

storing the spatial features of the first and second feature-extracted signals as binary values; and

determining the spatial feature match score by an exclusive OR operation between the first and second feature extracted signals, and

counting the number of logical one values in the result of the exclusive OR operation.

13. The method of claim 1 wherein the statistical attribute is either a mean, or a standard deviation, or both.

14. The method of claim 1 wherein the step of determining whether each of the first feature-extracted signal and second feature-extracted signal is a flat line signal further depends on, for each respective feature-extracted signal, a number of frames times a mean and a number of frames times a standard deviation, where number of frames is a count of the frames in the respective feature-extracted signal and where mean and standard deviation are statistical attributes of the respective feature-extracted signal.

15. A method for determining if a first video signal and second video signal match, comprising:

extracting a first set of mean pixel value features from the first video signal;

extracting a second set of mean pixel value features from the second video signal;

performing a first pass temporal correlation on the first and second mean pixel value features to provide a first pass temporal matching score;

if the first pass temporal matching score is not greater than a first pass threshold score, then

omitting the first and second video signals as a possible match;

else

if the first pass matching score is greater than a threshold score, then

performing a second pass, spatial frame comparison between the pixel value features of the first video signal and the second video signal, to provide a second pass spatial matching score;

if the second pass spatial matching score is not greater than a second pass threshold, then omitting the first and second video signals as a possible match; else

if the second pass spatial matching score is greater than the second pass threshold, then retaining the first and second video signals as a possible match.

16. The method of claim 15 wherein the pixel value feature extracted from the first and second video signals is a set of gray scale values, one gray scale value for each frame in each respective video signal.

19

17. The method of claim 15 further comprising:  
if the second pass spatial matching score is greater than the  
second pass threshold,  
determining a similarity score from the first pass temporal  
matching score. 5

18. The method of claim 15 further comprising:  
storing the pixel value features extracted from the first and  
second feature-extracted signals as binary values; and  
determining the match score from the spatial feature match  
using an exclusive OR operation and counting the num- 10  
ber of logical one values in the result.

\* \* \* \* \*

20

UNITED STATES PATENT AND TRADEMARK OFFICE  
**CERTIFICATE OF CORRECTION**

PATENT NO. : 9,087,125 B2  
APPLICATION NO. : 12/853008  
DATED : July 21, 2015  
INVENTOR(S) : Jared S. Berg

Page 1 of 1

It is certified that error appears in the above-identified patent and that said Letters Patent is hereby corrected as shown below:

In the Specification:

Col. 7, line 26 should read:

See FIG. 6 for an example flatline signal.

Col. 8, line 52 should read:

WAV file is created, or less than 5 seconds of WAV data

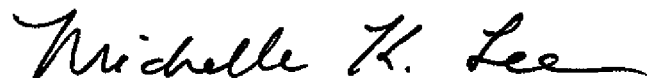
Col. 9, line 44 should read:

ways during the ingestion process. The first manner is

Col. 14, line 33 should read:

Network Interface Cards (NIC) 1306. The NIC(s) are con-

Signed and Sealed this  
Twenty-ninth Day of December, 2015



Michelle K. Lee  
*Director of the United States Patent and Trademark Office*